

*Semantic Wiki Conference, 24.-26. Nov. 2020*

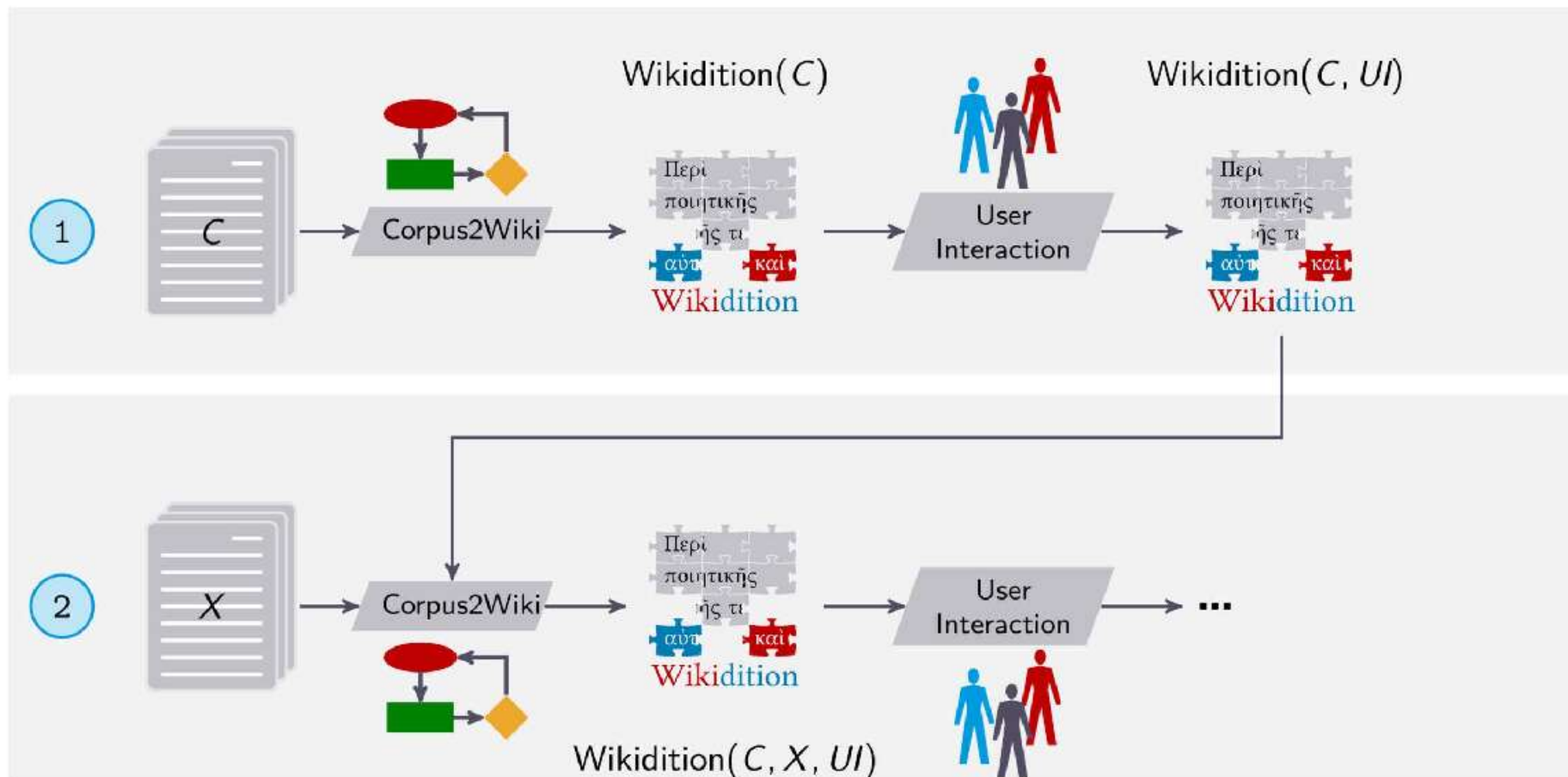
# Corpus2Wiki: A Tool for Automatically Generating Wikiditions in Digital Humanities

Alexander Mehler, Wahed Hemati  
Goethe-Universität Frankfurt




# Introduction

## From any Text Corpus to its Wikidition



# Introduction

## 1 Linkification: automatic multi-level, intra- & intertextual networking.

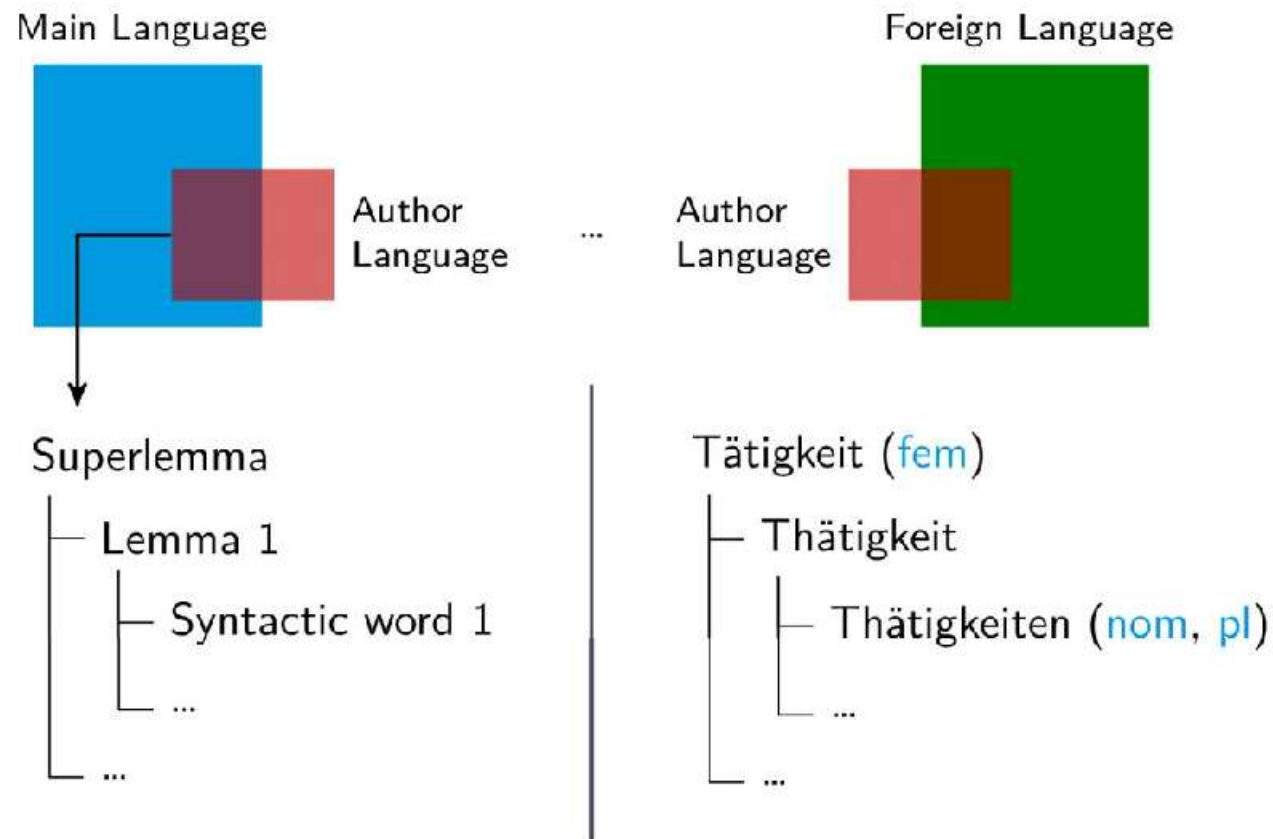
- 
1. Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte,
  2. fand er sich in seinem Bett zu einem ungeheuren Ungeziefer verwandelt.

1. Als er dies alles in größter Eile überlegte,
2. ohne sich entschließen zu können,
3. das Bett zu verlassen
4. – gerade schlug der Wecker dreiviertel sieben –
5. klopfte es vorsichtig an die Tür am Kopfende seines Bettes.

# Introduction

1 Linkification: automatic multi-level, intra- & intertextual networking.

2 Lexiconisation: *extracting corpus-specific lexica*.



# Introduction

- 1 Linkification: automatic multi-level, intra- & intertextual networking.
- 2 Lexiconisation: *extracting corpus-specific lexica*.
- 3 Little computational expertise required.
- 4 Interactivity and extensibility according to the Wiki Principle.



- 5 DH functions: curation, analysis, editing, modelling.

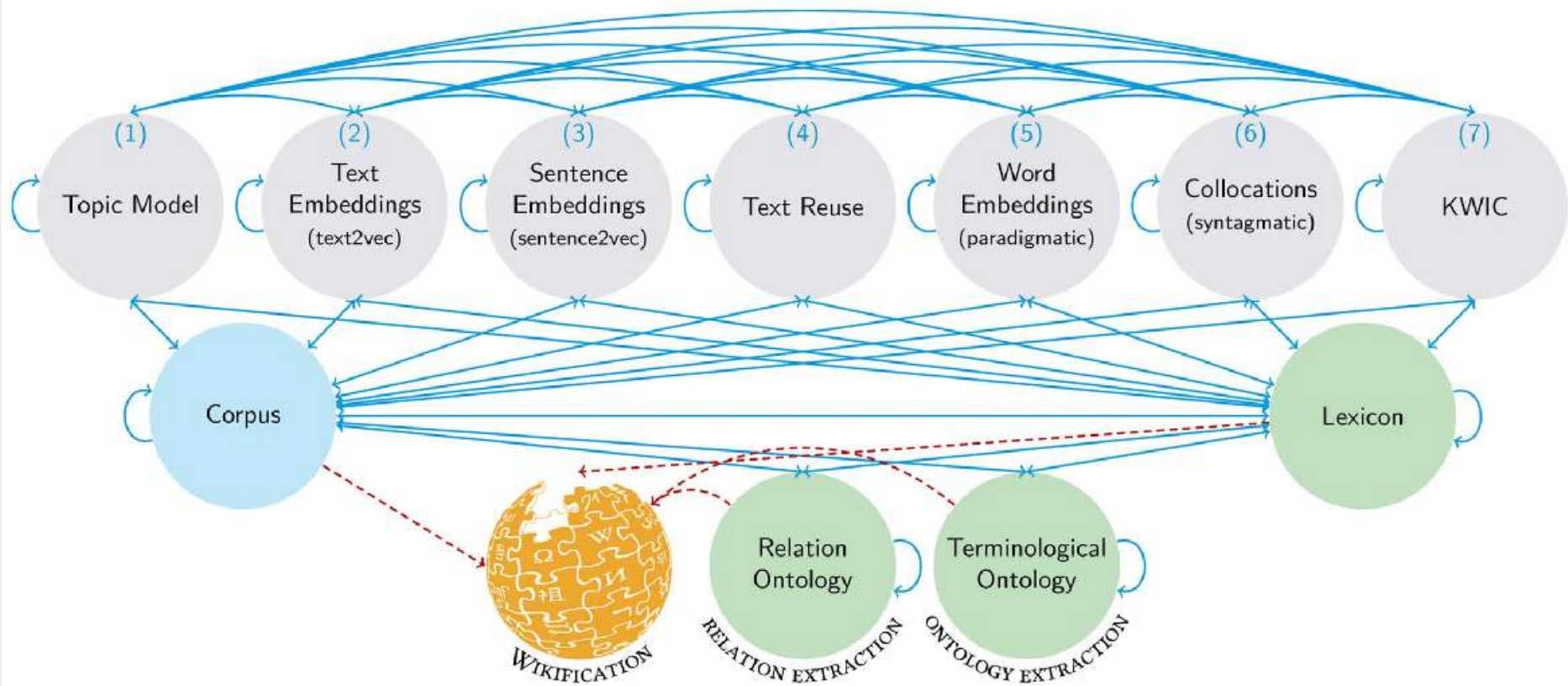
# Introduction

## Sign Model (Hjelmslev 1969)

Levels of Structure Formation					
		syntagmatic		paradigmatic	
		$a \leftarrow b$	$a \leftrightarrow b$	$a \leftarrow b$	$a \leftrightarrow b$
Word			+		+
Sentence			+		+
Text			+		+

Terminology:

1.  $a \leftarrow b$ : **determination** / selection / specification (relatedness)
2.  $a \leftrightarrow b$ : **interdependence** / solidarity / complementarity (similarity)



### + Transitivity:

- Each (lexical, sentential, ...) token on each article links to its type.
- Each (lexical, sentential, ...) type links to all its tokens.
- **Commuting diagram:**  
Traversing the network of texts, sentences and words in/from any direction/perspective.

### + Reconstructability: All computational models (vectors, ...) are part of the edition.

# Introduction

„Es ist ein eigentümlicher **Apparat**“, sagte der **Offizier** zu dem **Forschungsreisenden** [...].“  
(Franz Kafka (1919): In der Strafkolonie)

1. Heinrich Rauchberg - Statistische Technik

- » Ich weiss nicht,« sagte der Offizier, » ob Ihnen der Kommandant den Ap
- **Paradigmatic Similarity**
  1. » Ich kann nicht,« sagte der Reisende.
  2. » Lesen Sie,« sagte der Offizier.
  3. » Sie können es,« sagte der Offizier.
  4. [...]
- **Syntagmatic Similarity**
  1. » Es ist sehr kunstvoll,« sagte der Reisende ausweichend, » aber ich kann e
  2. » Ich kann nicht,« sagte der Reisende, » ich sagte schon, ich kann diese Blä
  3. » Ja gewiss,« sagte der Offizier lächelnd, » befühlen Sie es selbst.«
  4. [...]

2. sagen (V)

- **sagte** (sagen - V, Indicative, Singular, 3, Past)
- **Paradigmatic Similarity**
  1. bemerken (V)
  2. merken (V)
  3. meinen (V)
  4. [...]
- **Syntagmatic Relatedness (Collocations)**
  1. Offizier (NN)
  2. Reisender (NN)
  3. haben (V)
  4. [...]

Es kam abgeschlossen kleinen Tal ausser dem Offizier und dem Reisenden nur der Verurteilte, ein stumpfsinniger, at zugegen, der die schwere Kette hielt, in welche die kleinen Ketten ausliefen, mit denen der Verurteilte an den Fuss- und durch Verbindungsketten zusammenhingen.

schein hatte, als könnte man ihn frei auf den Abhängen herumlaufen lassen und müsse bei Beginn der Exekution nur

urteilten fast sichtbar unbeteiligt auf und ab, während der Offizier die letzten Vorbereitungen besorgte, bald unter den tief oberen Teile zu untersuchen.

lassen können, aber der Offizier führte sie mit einem grossen Eifer aus, sei es, dass er ein besonderer Anhänger dieses niemandem anvertrauen konnte.

ei zarte Damentaschentücher hinter den Uniformkragen gezwängt.

ende, statt sich, wie es der Offizier erwartet hatte, nach dem Apparat zu erkundigen.

nutzten Hände in einem bereitstehenden Wasserkübel, »aber sie bedeuten die Heimat; wir wollen nicht die Heimat rocknete die Hände mit einem Tuch und zeigte gleichzeitig auf den Apparat.

rat ganz allein.«

Es kommen natürlich Störungen vor; ich hoffe zwar, es wird heute keine eintreten, immerhin muss man mit ihnen rechnen. über auch Störungen vorkommen, so sind es doch nur ganz kleine und sie werden sofort behoben sein.«

laufen von Rohrstühlen einen hervor und bot ihn dem Reisenden an; dieser konnte nicht ablehnen.

rf.

aufgehäuft, zur anderen Seite stand der Apparat.

<sup>19</sup>»Ich weiss nicht,« sagte der Offizier, » ob Ihnen der Kommandant den Apparat schon erklärt hat.«

<sup>20</sup>Der Reisende machte eine ungewisse Handbewegung; der Offizier verlangte nichts Besseres, denn nun konnte er selbst den Apparat erklären.

<sup>21</sup>»Dieser Apparat,« sagte er und fasste eine Kurbelstange, auf die er sich stützte, »ist eine Erfindung unseres früheren Kommandanten. Ich habe gleich bei den allerersten Versuchen

# Agenda

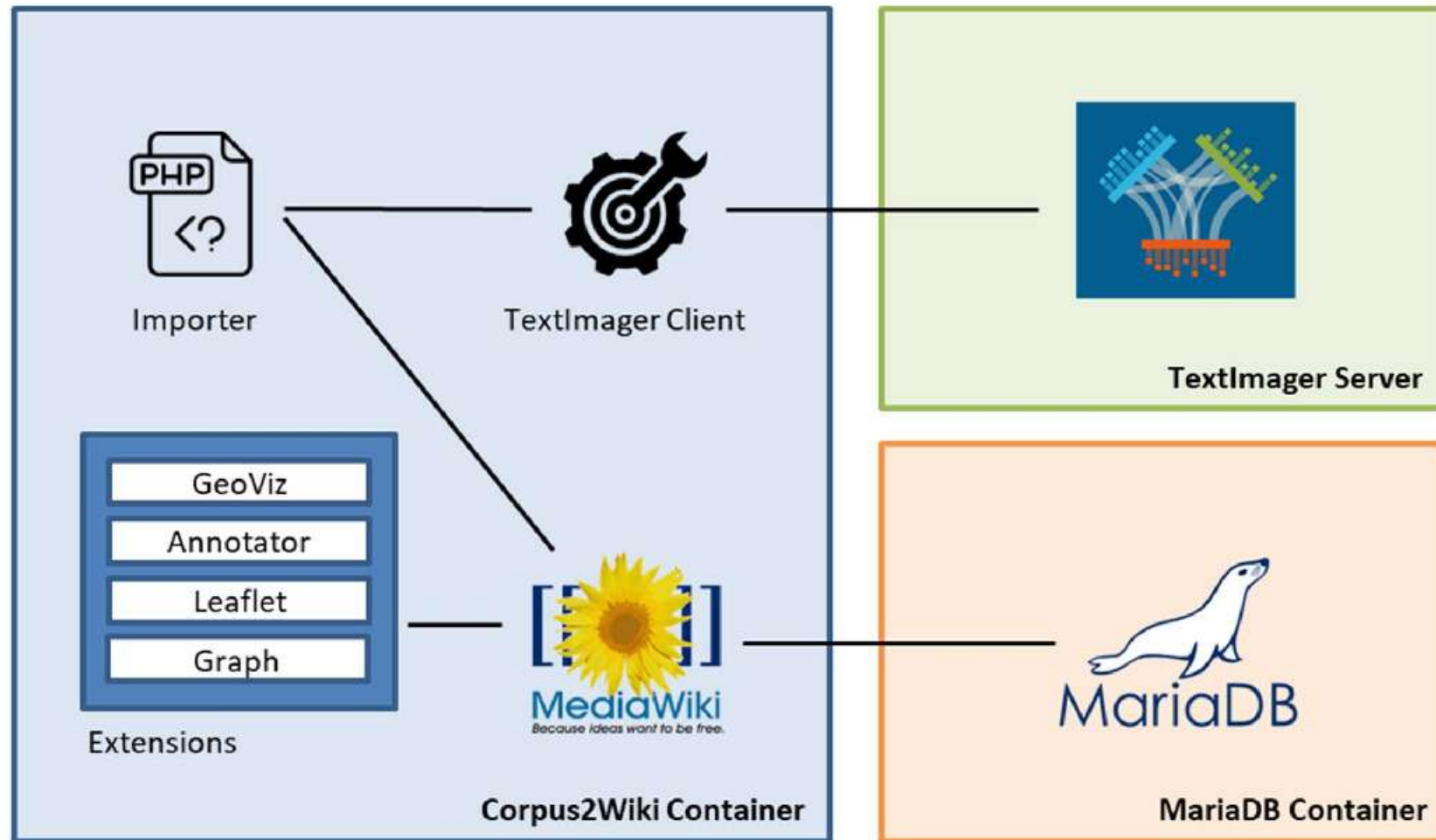
1. Introduction

2. Implementation

3. Conclusion

# Implementation

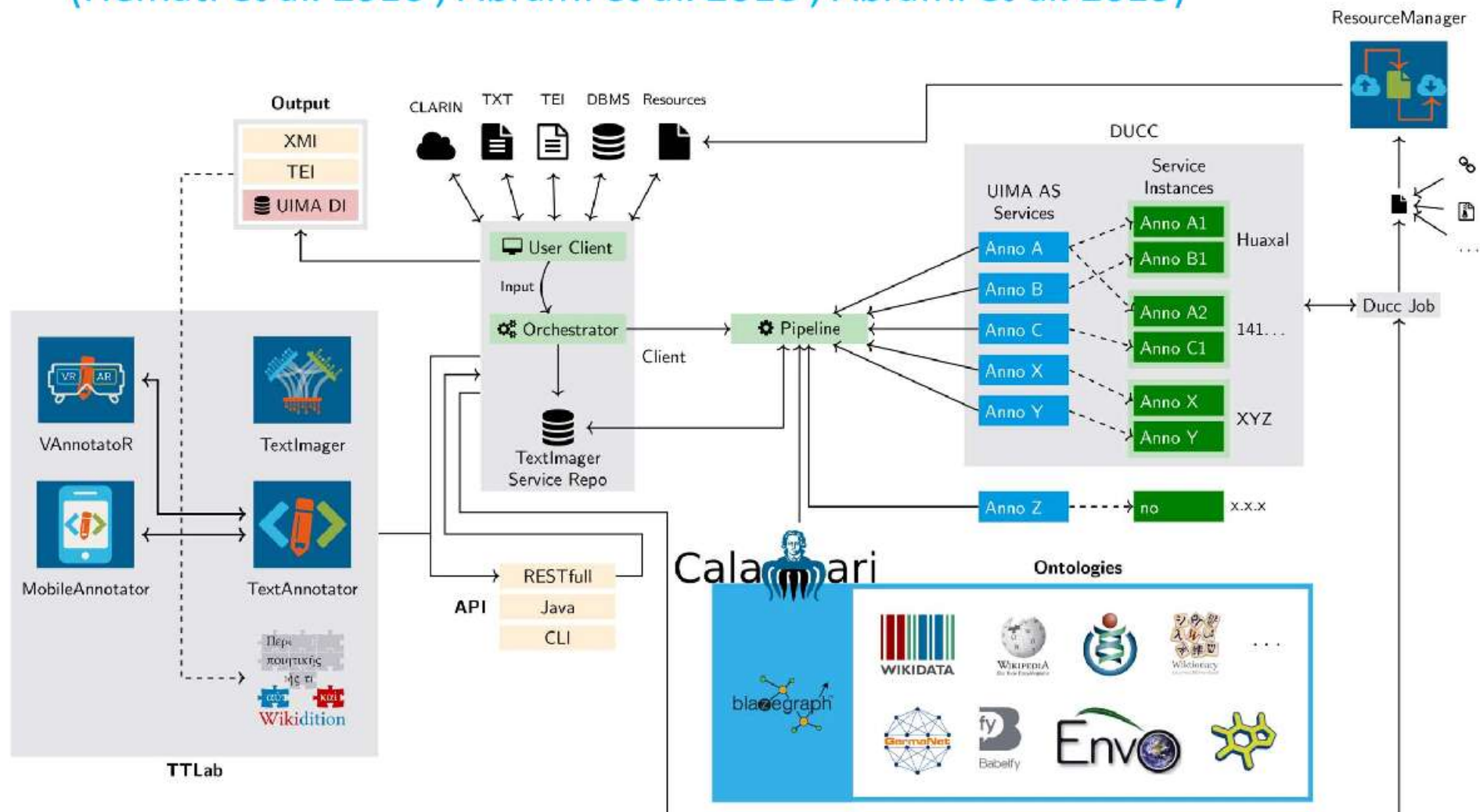
## Architecture (Hunziker et al. 2019)



## Implementation

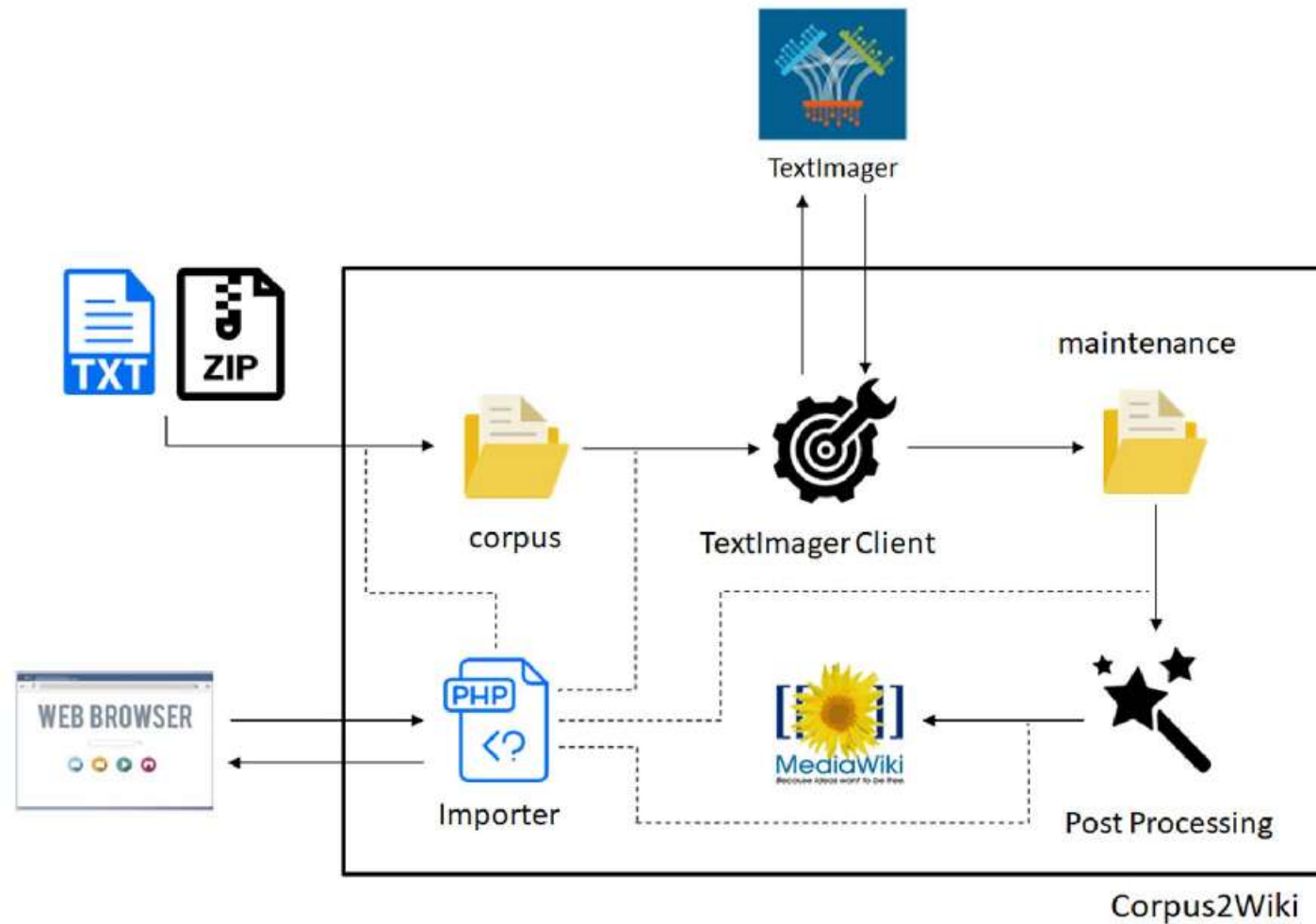
## Landscape: TextImager & TextAnnotator

(Hemati et al. 2016 ; Abrami et al. 2018 ; Abrami et al. 2019)



# Implementation

Import (Hunziker et al. 2019)



# Implementation

## Annotation & Visualization (Hunziker et al. 2019)

The screenshot displays a web application interface for text annotation and visualization. The main content area shows a text editor with the title "Johann Gottfried Herder-Kritische Waelder.txt". The text is annotated with various tags, including "Böhmer" and "witziger". A sidebar on the left contains navigation links such as "Home", "About", "Contact", and "Search". Below the text editor, there is a bar chart and a world map. A callout box highlights a specific text segment and its corresponding metadata.

**Text Information**

Field	Value
lemma	Böhmer
pos	NE
NE	I-PER

The callout box also displays the text segment: "Der Abt Böhmer und jene geistlichen, und ein witziger Mann, der Abt Trubler, mit einmal ermüdet von Scholiasten Schriften Erquickung sucht von furcht".

# Agenda

1. Introduction
2. Implementation
3. Conclusion

# Conclusion

## (Computational) Notions of Reading



	Human Close Reading	Distant Reading (Moretti 2013)	Machine Reading (Etzioni 2007)	Machine Close Reading
Research object	$\{T_1 \dots, T_m \mid X_n\}$ $m \rightarrow 1, n > 1$	$\{T_1, \dots, T_m\}$ $m \rightarrow \infty$	$\{T_1, \dots, T_m\}$ $m \rightarrow \infty$	$\{T_1 \dots, T_m \mid X_n\}$ $m \ll n, n \rightarrow \infty$
Quantity of data	small data	big data	small $\rightarrow$ big	big $\rightarrow$ small
Quantification	implicit	machine-based	machine-based	twofold
Interpretation	human-based	human-based	machine-based	human based
Research focus	understanding	hidden laws	understanding	hypotheses testing
Resources	human mind	corpus + HM	corpus + SW	corpus+HM+SW

- HM: Human Mind
- SW: Semantic Web

# Conclusion

## Application Scenarios (Wagner, Mehler, Biber 2016)

	$\{\text{hypotext}_i \mid i = 1\}$	$\{\text{hypotext}_i \mid i > 1\}$	$\{\text{hypotext}_i \mid i \gg 1\}$
$\{\text{hypertext}_i \mid i = 1\}$	(1)	(2)	(3)
$\{\text{hypertext}_i \mid i > 1\}$	(4)	(5)	(6)
$\{\text{hypertext}_i \mid i \gg 1\}$	(7)	(8)	(9)

Kafka's <i>Bericht für eine Akademie</i>	:	Hauff's <i>Der Affe als Mensch</i>	(1)
Kafka's <i>Bericht für eine Akademie</i>	:	all „Affentexte“ (Borgards 2012)	(2)
Kafka's <i>Beim Bau der ...</i>	:	Prager Tagblatt 1914-1917	(3)
Kafka's oeuvre	:	Nietzsche's <i>Geburt der Tragödie</i>	(4)
Kafka's oeuvre	:	Nietzsche's oeuvre	(5)
Kafka's Oeuvre	:	newspaper corpus	(6)
oeuvres of several authors	:	Goethe's <i>Faust</i>	(7)
oeuvres of several authors	:	a corpus of <i>Faust</i> texts	(8)
oeuvres of German authors	:	oeuvres of French authors	(9)

# References

- Abrami, Giuseppe und Alexander Mehler (2018). „A UIMA Database Interface for Managing NLP-related Text Annotations“. In: *Proceedings of the 11<sup>th</sup> edition of the Language Resources and Evaluation Conference, May 7 - 12*. LREC 2018. Miyazaki, Japan.
- Abrami, Giuseppe, Alexander Mehler, Andy Lücking, Elias Rieb und Philipp Helfrich (Mai 2019). „TextAnnotator: A flexible framework for semantic annotations“. In: *Proceedings of the Fifteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation, (ISA-15)*. ISA-15. Gothenburg, Sweden.
- Borgards, Roland (2012). „Tiere in der Literatur. Eine methodische Standortbestimmung“. In: *Das Tier an sich: Disziplinenübergreifende Perspektiven für neue Wege im wissenschaftsbasierten Tierschutz*. Hrsg. von Herwig Grimm und Carola Otterstedt, S. 87–118.
- Etzioni, Oren (2007). „Machine Reading of Web Text“. In: *Proceedings of the 4th International Conference on Knowledge Capture*. K-CAP '07. Whistler, BC, Canada, S. 1–4.
- Hemati, Wahed, Tolga Uslu und Alexander Mehler (2016). „TextImager: a Distributed UIMA-based System for NLP“. In: *Proc. of COLING 2016: System Demonstrations*. Osaka, Japan, S. 59–63.
- Hjelmslev, Louis (1969). *Prolegomena to a Theory of Language*. Madison: University of Wisconsin Press.
- Hunziker, Alex, Hasanagha Mammadov, Wahed Hemati und Alexander Mehler (2019). „Corpus2Wiki: A MediaWiki-based Tool for Automatically Generating Wikiditions in Digital Humanities“. In: *INF-DH-2019*. Hrsg. von Manuel Burghardt und Claudia Müller-Birn. Bonn: Gesellschaft für Informatik e.V.
- Mehler, Alexander, Rüdiger Gleim u. a. (2016). „Wikidition: Automatic Lexiconization and Linkification of Text Corpora“. In: *Information Technology* 58.2, S. 70–79.
- Mehler, Alexander, Benno Wagner und Rüdiger Gleim (2016). „Wikidition: Towards A Multi-layer Network Model of Intertextuality“. In: *Proceedings of DH 2016, 12-16 July*. DH 2016. Kraków.
- Moretti, Franco (2013). *Distant Reading*. Verso.
- Rutherford, Eleanor, Wahed Hemati und Alexander Mehler (2018). „Corpus2Wiki: A MediaWiki based Annotation & Visualisation Tool for the Digital Humanities“. In: *INF-DH-2018*. Hrsg. von Manuel Burghardt und Claudia Müller-Birn. Bonn: Gesellschaft für Informatik e.V.
- Wagner, Benno, Alexander Mehler und Hanno Biber (2016). „Transbiblionome Daten in der Literaturwissenschaft. Texttechnologische Erschließung und digitale Visualisierung intertextueller Beziehungen digitaler Korpora“. In: *DHd 2016*.